Повесьте ваши уши на гвоздь внимания !!!!!!

Элементы математической статистики.

Случайные выборки. Первичная обработка статистических данных. Вариационные ряды.

Статистика изучает большие массивы информации и устанавливает закономерности, которым подчиняются случайные массовые явления.

Генеральной совокупностью (ГС) называется вся подлежащая изучению какого-либо свойства (говорят, признака) совокупность объектов.

Та часть объектов, которая отобрана для непосредственного изучения какого-либо признака ГС носит название **случайной выборки** (или просто **выборки**).

Объем ГС и объем выборки – это количество элементов в них. Обозначаются , соответственно, N и n.

В дальнейшем будем считать, что объем выборки существенно меньше объема генеральной совокупности. В этом случае получаемые в дальнейшем формулы являются наиболее простыми.

Непрерывная природа изучаемого признака порождает бесконечные ГС.

Для того, чтобы выборка была **репрезентативной** (хорошо представлять элементы ГС), она должна быть отобрана случайно. Случайность отбора элементов в выборку достигается соблюдением принципа равной возможности каждого элемента ГС быть отобранным в выборку.

Нарушение принципов случайного выбора приводит к серьезным ошибкам.

Любое число, полученное на основе выборки, носит название **«выборочная статистика»** (или просто «статистика»).

Пусть получена выборка объема n. Над этим массивом исходных данных выполняется операция ранжирования, т.е. экспериментальные данные выстраиваются в порядке возрастания:

$x_1 < x_2 < x_3 < \ldots < x_k; \quad k \le n;$

причем значение x_i *встречается* n_i *раз*:

$n_1 + n_2 + \ldots + n_k = n;$

вводится терминология:

x_i – *вариант*; n_i – *частота варианта*

(*количество появлений значений* x_i);

$w_i = \frac{n_i}{n}$ – *относительная частота варианта или частость*;

обязательно выполняется $\sum_{i=1}^{k} w_i = 1;$

размах выборки $R = x_{max} - x_{\min} = x_k - x_1.$

Определение.

Вариационным рядом называется ранжированный в порядке возрастания ряд значений (вариантов) с соответствующими им частотами.

Значения x_i	x_1	x_2	...	x_k
Частоты n_i	n_1	n_2	...	n_k
Частости $w_i=n_i/n$	w_1	w_2	... ...	w_k

Данный вариационный ряд носит название дискретного вариационного ряда (его члены принимают отдельные изолированные значения).

Построение дискретного вариационного ряда нецелесообразно, когда число значений в выборке велико или признак имеет непрерывную природу, т.е. может принимать любые значения в пределах некоторого интервала. В этом случае строят интервальный вариационный ряд.

Вид интервального ряда:

Интервалы вариантов	$x_1 - x_2$ *1*	$x_2 - x_3$ *2*	...	$x_{k-1} - x_k$ *k-1*
Частоты n_i *(число вар-тов, попавших в инт-вал)*	n_1	n_2	...	n_{k-1}
Частости $w_i = n_i/n$	w_1	w_2	...	w_{k-1}

В том случае, когда можно предположить, что изучаемый признак в ГС подчиняются нормальному з.р., для вычисления количества интервалов равной длины применяют формулу Стерджесса:

$$m = 1 + 3.3 \cdot \lg n, \quad \text{если} \quad m \in [6;12]$$

Если $m > 12$, то принимают $m = 12$;

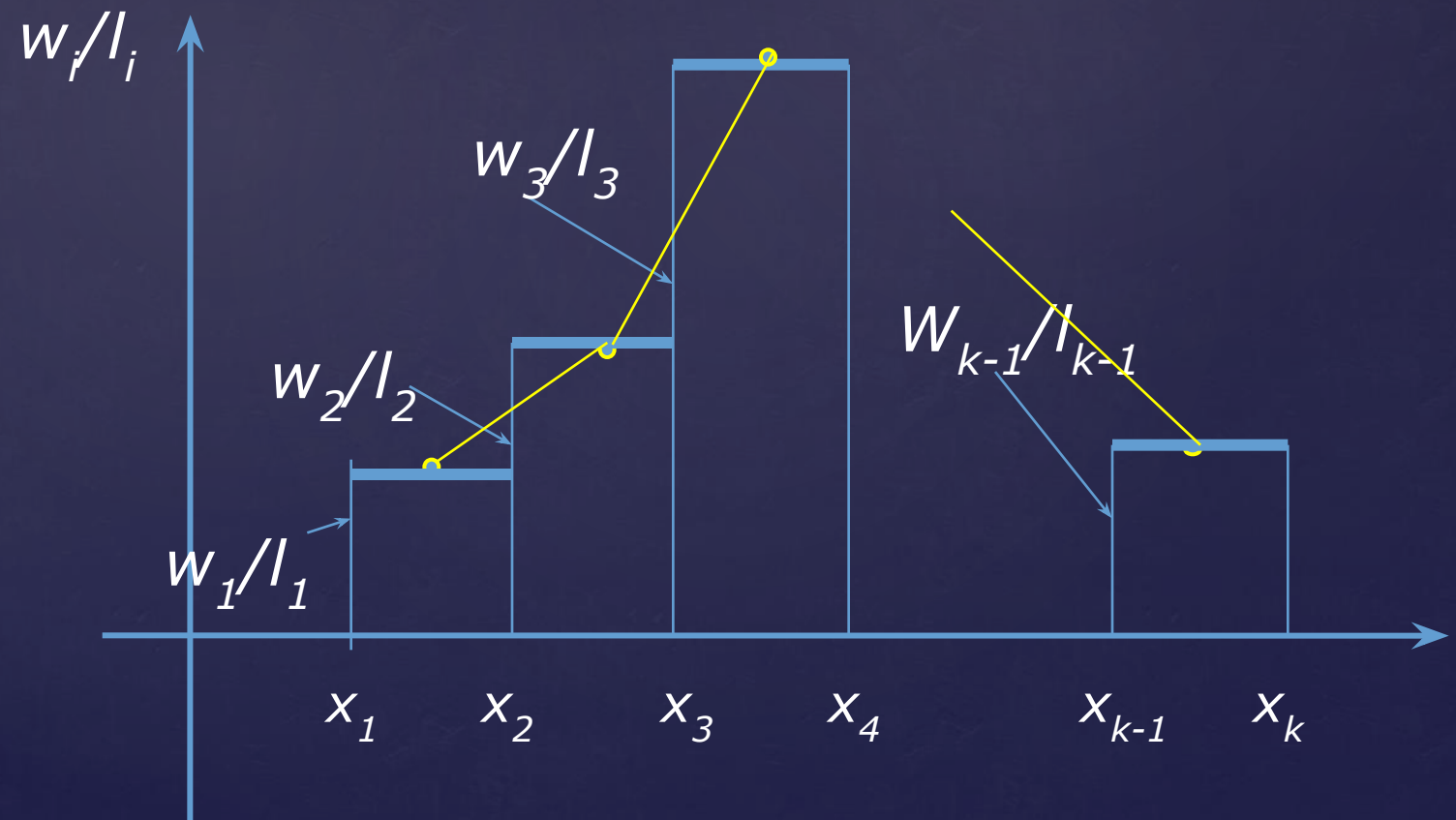
если $m < 6$, то принимают $m = 6$.

Длина отдельного интервала: $l = \frac{R}{m} = \frac{x_{\max} - x_{\min}}{m}$

Существуют различные **приёмы изображения** набора данных, которые дают визуальное представление об основных свойствах экспериментальных данных в целом. Чаще всего для этого используются: полигон, гистограмма, кумулята. Графическое представление вариационных рядов делает картину поведения статистических данных более наглядной.

Полигон распределения частот используется для изображения дискретного вариационного ряда и представляет собой ломаную линию, отрезки которой соединяют точки с координатами (x_i, w_i).

Гистограмма используется для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру из прямоугольников с основаниями, равными интервалам значений признака l_i ($l_i = x_{i+1} - x_i$) и высотами, равными w_i/l_i.



Эмпирической функцией распределения $F_n(x)$ называется относительная частота того, что случайная величина принимает значение меньше заданного:

$$F_n(x) = W(X<x) = W_x^{нак}$$

Для графического изображения эмпирической функции распределения служит кумулята. Строим ее, соединяя точки $(x_i, W_i^{нак})$.

Следует дополнить вариационные ряды и их графическое изображение некоторыми сводными характеристиками вариационных рядов.

Эти обобщающие показатели в компактном виде характеризуют всю выборку (вариационный ряд) в целом. К таким обобщающим показателям относят:

)Характеристики центральной тенденции - это средние величины, определяющие значения признака, вокруг которого концентрируются все его наблюдаемые значения;

)Характеристики вариации (изменчивости) – это величины, определяющие колебания наблюдаемых значений признака.

В качестве основной характеристики центральной тенденции чаще всего используют среднее арифметическое, вычисленной на основе выборки. Помимо этой величины используют моду и медиану.

Определение:

Медиана – это значение признака, приходящееся на середину ранжированного ряда наблюдений.
Иначе: это то значение варианта, которое делит вариационный ряд на две равные по объему части.

Обозначение:
Теоретическое $MeX;$

Статистическое $\tilde{Me}$ Me

Если число вариант нечетное, т.е. n=2m+1 , то $\tilde{Me} = x_{m+1}$

Если число вариант четное, т.е. n=2m , то $\tilde{Me} = (x_m + x_{m+1})/2$

Определение:

Модой называется значение признака, наиболее часто встречающееся в выборке.

Иначе:

Мода - то значение варианта, которому соответствует наибольшая частота.

Обозначение:

Теоретическое Mo_X;

Статистическое $\tilde{Mo}$

Нам важно знать не только средние значения вариантов, но и отличие значений вариантов от среднего значения. Для отражения изменчивости (вариации) значений признака вводят различные показатели вариации ряда.

Простейшим и весьма приближенным показателем вариации является размах выборки $R = x_{max} - x_{min}$.

Определение.

Выборочной дисперсией вариационного ряда называется среднее арифметическое квадратов отклонений вариантов от их среднего арифметического:

$$S_*^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n} = \sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 \cdot w_i$$

При вычислении выборочной (или эмпирической) дисперсии формулу несколько меняют. Из некоторых соображений, которые пока для нас с вами скрыты, в знаменателе этой формулы ставят не *n*, а *n-1*, и возникает другая формула для вычисления дисперсии, которую запишем ниже; величину, вычисленную по этой формуле называют «исправленная выборочная дисперсия».

$$S^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n-1} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n-1}$$

Будем всегда выборочную дисперсию вычислять по второй формулу, называя ее просто «выборочная дисперсия». Ясно, что при большом объеме выборки разница между двумя приведенными формулами стирается.

Для меры вариации, выраженной в тех же единицах измерения, что и значение признака, вычисляют выборочное стандартное отклонение:

$$S = \sqrt{S^2} = \sqrt{\sum_{i=1}^{k} \frac{\left(x_i - \bar{x}\right)^2 \cdot n_i}{n-1}}$$

Для сравнения вариаций разных по природе переменных используется относительный показатель вариации:

$$\textit{Коэффициент вариации} \quad V = \frac{S}{\bar{x}} \cdot 100\%$$

Эта величина характеризует, насколько сильно элементы в выборке и, следовательно, в ГС отличаются друг от друга.

Точечные оценки параметров генеральной совокупности.

Поставим задачу в общем виде – задачу отыскания хороших (доброкачественных) приближений параметров известных распределений на основе выборки из ГС.

Пусть $x_1, x_2, \ldots, x_n$ - выборка объема n из ГС. Будем рассматривать эту выборку как систему СВ $X_1, X_2, \ldots, X_n$, которая в данном конкретном исследовании приняла именно этот набор числовых значений $x_1, x_2, \ldots, x_n$.

Определение:

Точечной оценкой $\tilde{\theta}_n$ неизвестного параметра θ теоретического закона распределения называют всякую функцию результатов наблюдений над СВ X, значение которой принимают в качестве приближённых значений параметра θ :

$$\tilde{\theta}_n = f(x_1, x_2, \ldots, x_n)$$

Требования, предъявляемые к точечным оценкам
(Иногда говорят : *свойства точечных оценок):*

1.Несмещённость.

Оценка $\tilde{\theta}_n$ параметра θ называется **несмещённой**, если её математическое ожидание равно оцениваемому параметру:

$$E\tilde{\theta}_n = \theta$$

2. Эффективность.

Оценка $\tilde{\theta}_n$ параметра θ называется **эффективной**, если она имеет наименьшую дисперсию среди всех оценок параметра по выборкам одного и того же объема:

$$D\tilde{\theta}_n \to \min \quad при \quad фиксир. \quad значении \quad n.$$

3. Состоятельность.

Оценка $\tilde{\theta}_n$ параметра θ называется **состоятельной**, если она удовлетворяет ЗБЧ:

$$\tilde{\theta}_n \xrightarrow[n\to\infty]{P} \theta$$

В последнее время стали добавлять еще одно требование к оценкам.

4. Устойчивость.

Смысл этого свойства в том, что при небольших флуктуациях в исходной информации значение оценки не должно существенным образом меняться.

На практике не всегда удается удовлетворить всем требованиям одновременно. *Может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенные оценки или же оценки, обладающие несколько большей дисперсией по сравнению с эффективными оценками.*

Показано, что среднее арифметическое, вычисленное на основе выборки и являющееся точечной оценкой генерального среднего (истинного значения параметра), обладает свойствами 1-4, присущими хорошей оценке.

Показано также, что **выборочная доля *w=k/n*** (иначе: относительная частота появления признака в выборке) является несмещенной и состоятельной оценкой генеральной доли $\mathbf{W_г=K/N}$.

Заметим, что выборочную долю можно трактовать как оценку вероятности в биномиальном законе распределения.

Показано, что выборочная дисперсия, вычисляемая по формуле

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1},$$

дает несмещенную оценку генеральной дисперсии.

Аналогично, несмещенной точечной оценкой ковариации *cov(X,Y)* является такая оценка:

$$K_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right) - \textit{выборочная ковариация.}$$

В формулах для S^2 и K_{XY} возникает новый параметр **k=n-1**

Он носит название *«число степеней свободы».*Это разность между числом используемых в расчетах отклонений и количеством связей между этими отклонениями.

Методы получения точечных оценок параметров генеральной совокупности.

Основное внимание уделим методу, который наиболее часто применяется для этой цели.

1. Метод наибольшего (максимального) правдоподобия.

- это основной метод получения оценок параметров ГС на основе выборки. Метод был предложен американским статистиком Р. Фишером.

Пусть задан известный закон распределения. Ставится задача найти оценку его неизвестного параметра или параметров, если в законе распределения их несколько.

Функцией правдоподобия дискретной СВ X называют функцию аргумента θ (искомого параметра)

$$L(x_1, x_2, \ldots, x_n; \theta) = p(x_1; \theta) \cdot p(x_2; \theta) \cdot \ldots \cdot p(x_n; \theta), \quad где$$

$$x_1, x_2, \ldots, x_n - фиксированные \quad числа.$$

В качестве точечной оценки параметра θ принимают такое его значение $\tilde{\theta}_n$, при котором функция правдоподобия достигает максимума. Оценку $\tilde{\theta}_n$ называют оценкой наибольшего правдоподобия.

Суть подхода заключается в том, чтобы выбрать такое значение оценки параметра, которое обеспечивает наиболее вероятное появление именно данной выборки.

Удобнее рассматривать не саму функцию *L*, а *lnL.*

Методом наибольшего правдоподобия найдена оценка параметра λ в законе распределения Пуассона

$$P(X=k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Методом наибольшего правдоподобия найдена оценка вероятности успеха в единичном испытании на основе единственной серии испытаний.

Методом наибольшего правдоподобия найдена оценка вероятности успеха в единичном испытании на основе нескольких серий испытаний (биномиальный закон распределения).

Функцией правдоподобия непрерывной СВ X называют функцию аргумента θ (искомого параметра)

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Здесь x_1, x_2, …, x_n - фиксированные числа.

Методом наибольшего правдоподобия найдена оценка параметра λ показательного з.р.

Методом наибольшего правдоподобия найти оценки параметров m и σ нормального з.р.

По поводу метода наибольшего правдоподобия сделаем **выводы:**

1. Метод наибольшего правдоподобия дает естественные оценки, не противоречащие здравому смыслу.

Усилиями математиков было показано, что в целом эти оценки обладают хорошими свойствам. А именно, они являются состоятельными, эффективными, но иногда слабо смещенными.

2. Метод наибольшего правдоподобия имеет два недостатка:

1) иногда сложно решить уравнение или систему уравнений правдоподобия, которые часто бывают нелинейными.

2) существенное ограничение метода – необходимо точно знать вид закона распределения, что во многих случаях оказывается невозможным.

Существует и другие методы нахождения точечных оценок параметров ГС. Это – **Метод моментов** и

Метод наименьших квадратов.

Суть его заключается в том, что оценка определяется из условия минимизации квадратов отклонений выборочных данных от определяемой оценки.

Следует ввести дополнительные распределения и новые таблицы, созданные на основе этих распределений.

Распределения, связанные с нормальным законом распределения.

.Распределение X - квадрат (X^2).

(или распределение Пирсона)

Определение:

Пусть СВ X_1, X_2, …, X_k независимые и каждая из них имеет стандартное нормальное распределение

($X_i \sim N(0;1)$, $i=1, 2,\ldots, n$), тогда случайная величина

$$X^2(k) = X_1^2 + X_2^2 + \ldots + X_k^2$$

имеет распределение хи-квадрат с k степенями свободы.

Значения этого распределения затабулированы.

2. t -распределение *(или распределение Стьюдента)*

Определение:

Пусть СВ Y, X_1, X_2, …, X_k независимые и каждая из них имеет стандартное нормальное распределение

$(Y, X_i \sim N(0;1), \quad i=1, 2,\ldots, k)$,

тогда случайная величина

$$t(k) = \frac{Y}{\sqrt{\frac{1}{k}\left(X_1^2 + X_2^2 + \ldots + X_k^2\right)}} = \frac{Y}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} X_i^2}} = \frac{Y}{\sqrt{\frac{1}{k}\chi^2(k)}}$$

имеет распределение Стьюдента с k степенями свободы.

Значения распределения затабулированы.

Интервальные оценки параметров генеральной совокупности.

Наша задача - научиться отыскивать границы интервала, который накроет истинное значение искомого параметра. Для этого будем использовать метод интервального оценивания, который разработал американский статистик Нейман, исходя из идей статистика Фишера. Этот интервал должен накрывать истинное значение параметра θ с большой вероятностью $\gamma = 1-a$, где γ - велико, а a - мало;

γ называется доверительной вероятностью (а также: надежностью, уровнем доверия), a называется уровнем значимости.

Интервал, который мы будем находить, носит название доверительного интервала (иначе: интервальная оценка искомого параметра ГС).

Ставится задача отыскания такого значения ε, для которого выполнено:

$$P\left(\left|\theta-\tilde{\theta}\right|<\varepsilon\right)=P\left(\underbrace{\tilde{\theta}-\varepsilon}_{\tilde{\theta}_1}<\theta<\underbrace{\tilde{\theta}+\varepsilon}_{\tilde{\theta}_2}\right)=\gamma$$

$\tilde{\theta}_1=\tilde{\theta}-\varepsilon;\quad \tilde{\theta}_2=\tilde{\theta}+\varepsilon$ –

– границы доверительного интервала;

$I_\gamma=\left(\tilde{\theta}_1;\tilde{\theta}_2\right)$ – доверительный интервал.

Величина ε называется **«точность оценки»** (или: **«предельная ошибка выборки»**).

Формулы, по которым определяются границы доверительного интервала, зависят от конкретного оцениваемого параметра ГС и конкретной ситуации, поэтому возникает необходимость рассмотреть несколько интересующих нас ситуаций.

1. **Интервальная оценка математического ожидания (или: генерального среднего) нормально распределенной ГС, если известна дисперсия σ^2 для ГС.**

Пусть изучаемый признак X в ГС имеет нормальное распределение с параметрами m и σ независимых СВ. В данной постановке задачи считаем, что σ^2 известна (например, взята из аналогичного предыдущего исследования).

Здесь m – тот неизвестный параметр, для которого мы хотим построить интервальную оценку.

Получено следующее выражение для доверительного интервала:

$$I_\gamma = \left(\overline{x} - \varepsilon; \overline{x} + \varepsilon\right), \quad где \quad \varepsilon = \frac{t_{кр} \cdot \sigma}{\sqrt{n}} = t_{кр} \cdot \sigma_{\overline{X}}$$

(С помощью таблицы функции Φ_0 находим по заданному значению γ $t_{кр}$ - квантиль стандартного нормального з.р. на основе уравнения $\Phi_0(t_{кр}) = \gamma/2$).

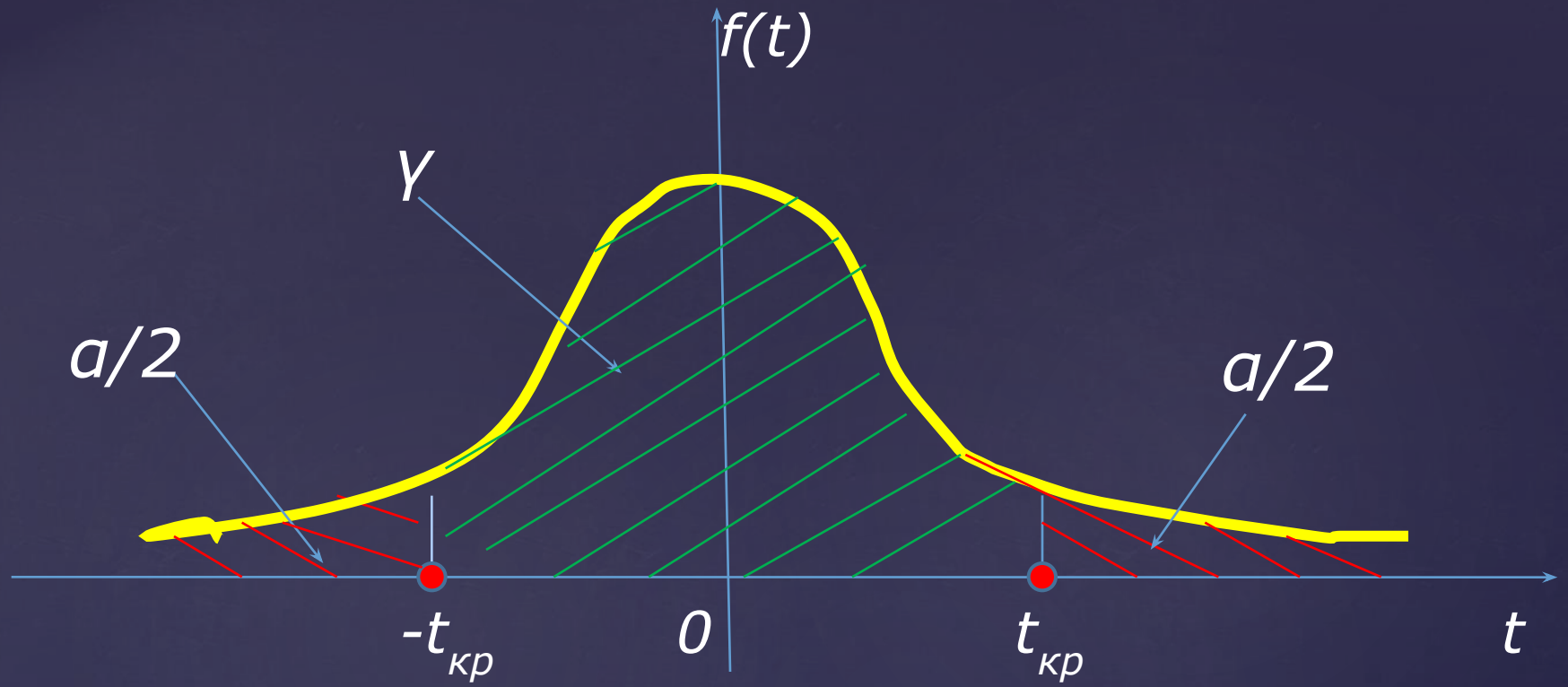
2. Интервальная оценка математического ожидания нормально распределенной ГС, если дисперсия σ^2 для ГС неизвестна.

Теперь вместо неизвестной дисперсии будем использовать ее точечную оценку – выборочную дисперсию

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$$I_\gamma = \left(\overline{x} - \varepsilon; \overline{x} + \varepsilon\right), \quad где \quad \varepsilon = \frac{t_{кр} \cdot S}{\sqrt{n}} = t_{кр} \cdot S_{\overline{X}}$$

(С помощью таблица **«Критические точки распределения Стьюдента»** по заданным значениям a (двусторонняя критическая область) и $k=n-1$ находим $t_{кр}$ - квантиль распределения Стьюдента).



Замечание:

При $n \leq 30$ (**малые выборки**) следует находить $t_{кр}$ на основе распределения Стьюдента;

При $n > 30$ (**большие выборки**) следует находить $t_{кр}$ на основе стандартного нормального распределения, т.е. на основе функции Лапласа.

Если задана точность оценки ε , то можно найти объем выборки, которая обеспечит эту требуемую точность:

$$n_{\min} = \left(\frac{t_{кр} \cdot S}{\varepsilon} \right)^2; \quad при \quad n \geq n_{\min} \quad эта$$

$$точность \quad будет \quad обеспечена.$$

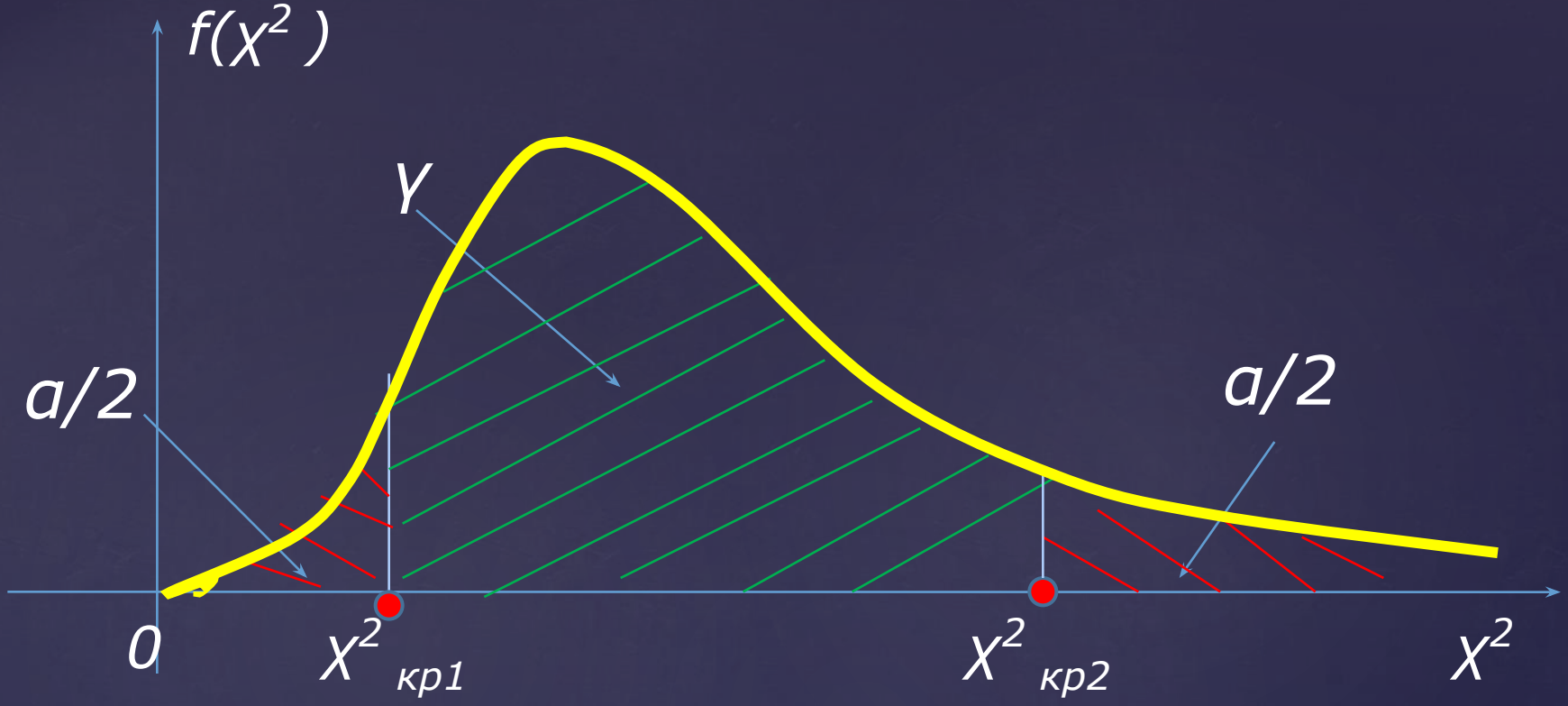
3. Интервальная оценка стандартного отклонения для нормально распределенной ГС.

Пусть изучаемый признак X в ГС имеет нормальное распределение: $X \sim N(m,\sigma)$, причем параметры распределения неизвестны.

Для случая малых объемов выборки ($n \leq 30$):

$$Доверительный \quad интервал \quad для \quad \sigma \quad имеет \quad вид:$$

$$I_\gamma = \left(S \cdot \sqrt{\frac{(n-1)}{\chi^2_{кр_2}}}; \quad S \cdot \sqrt{\frac{(n-1)}{\chi^2_{кр_1}}} \right)$$



Очевидно, что значения $X^2_{кр1}$ и $X^2_{кр2}$ определяются неоднозначно при одном и том же значении заштрихованной площади, равной γ . Границы красных зон выбираем так, чтобы вероятности попадания в них были бы одинаковыми, равными $a/2$.

Для случая больших объемов выборки (*n>30):*

$$I_\gamma = \left(S \cdot \frac{\sqrt{2(n-1)}}{\sqrt{2 \cdot n - 3} + t_{кр}}; \quad S \cdot \frac{\sqrt{2(n-1)}}{\sqrt{2 \cdot n - 3} - t_{кр}} \right),$$

где $t_{кр}$ *находим из табл. решения уравнения:* $\Phi_0(t_{кр}) = \frac{\gamma}{2}.$

4. Интервальная оценка истинного значения вероятности биномиального закона распределения (генеральной доли).

Рассмотрим два случая:

А. Случай умеренно больших выборок
(n>30 до нескольких сотен, например, до 200).

Далее в формуле $t_{кр}$ - квантиль стандартного нормального з.р. на основе уравнения $\Phi_0(t_{кр}) = \gamma /2.$

$$p_{1,2} = \frac{\left(w + \frac{t_{кр}^{\,2}}{2n}\right) \pm t_{кр} \cdot \sqrt{\frac{w(1-w)}{n} + \frac{t_{кр}^{\,2}}{4n^2}}}{\left(w + \frac{t_{кр}^{\,2}}{n}\right)} =$$

$$= \frac{n}{\left(n + t_{кр}^{\,2}\right)} \cdot \left(w + \frac{t_{кр}^{\,2}}{2n} \pm t_{кр} \cdot \sqrt{\frac{w(1-w)}{n} + \frac{t_{кр}^{\,2}}{4n^2}}\right) \rightarrow$$

$\rightarrow$ доверит. интервал для p находится следующим образом: $p_1 < p < p_2$,

где p_1, p_2 – меньший и больший корни этого уравнения; иначе: $I_\gamma = (p_1; \quad p_2)$.

Б. Случай больших выборок

(порядка сотен и более ; например, от 200 и более).

Формулы для вычисления границ доверительного интервала существенно упрощаются при таких больших объемах выборок.

$$I_\gamma = \left(w - \varepsilon; \quad w + \varepsilon \right), где \quad \varepsilon = t_{кр} \cdot \sqrt{\frac{w(1-w)}{n}} = t_{кр} \cdot S_w$$

При больших объемах выборок n возникает простая формула для ε, на основе которой при заданном ε можно вычислить соответствующее n:

$$n_{\min} = \frac{t_{кр}^2 \cdot w \cdot (1-w)}{\varepsilon^2}.$$

В. Случай выборок малого объема (*n≤30*)

В этом случае для вычисления S_w используется формула

$$S_w = \sqrt{\frac{w \cdot (1 - w)}{n - 1}}$$

Доверительный интервал определяется по формуле предыдущего пункта; $t_{кр}$ находится по распределению Стьюдента по *к=n-1.*

Замечание:

В литературе часто приводят упрощенный способ вычисления доверительного интервала, рассматривая только большие и малые выборки. В этом случае выделяют два пункта при вычислении доверительного интервала:

1) Большая выборка (*n* более 30) - вычисление ведут по пункту Б.
2) Малая выборка (*n* меньше или равно 30) – вычисление ведут по пункту В.

Благодарю за внимание!